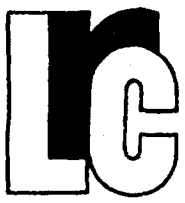
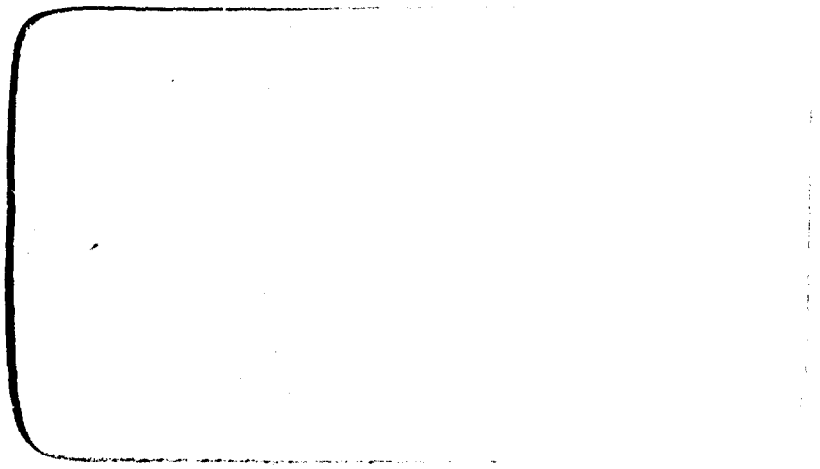




THIS IS A WORKING PAPER ☐ IT MAY BE EXPANDED, MODIFIED OR WITHDRAWN AT ANY TIME ☐ THE VIEWS, CONCLUSIONS, AND RECOMMENDATIONS EXPRESSED HEREIN DO NOT NECESSARILY REFLECT THE OFFICIAL VIEWS OF THE SPONSOR ☐

LINGUISTICS RESEARCH CENTER

THE UNIVERSITY OF TEXAS

BOX 7814 • UNIVERSITY STATION • AUSTIN 12, TEXAS

RETRIEVAL SYSTEM EXPERIMENTATION AND EVALUATION AT LRC

A. G. Dale

prepared for
National Science Foundation
Grant NSF GN-308

LINGUISTICS RESEARCH CENTER
The University of Texas
Box 7247, University Station
Austin, Texas 78712

LRC 65 WT-2

December 1965

CONTENTS

ABSTRACT

A time-shared experimental document retrieval system under development at The University of Texas is briefly described. A method for evaluating the effect on retrieval performance of controlled changes in the retrieval processor is proposed.

FOREWORD

This prototype time-shared system is being implemented by the Linguistics Research Center in cooperation with the Computation Center of The University of Texas. Preliminary work on the system was supported by the National Science Foundation under grant GN-308 and by the United States Army Electronics Laboratories under contract DA 36-039 AMC-02162 (E). Part of the equipment cost has been contributed by Control Data Corporation and part by the Excellence Fund of The University of Texas. Computation Center personnel have worked on the project under National Science Foundation grant GU-1010 to The University of Texas.

The paper was given at the NATO Advanced Study Institute on Evaluation of Information Retrieval Systems, The Hague, July 12-23, 1965.

1 INTRODUCTION

Research in information retrieval at the Linguistics Research Center is concerned with computer-based systems of a special type, namely interactive time-shared systems in which the requester is communicating directly with a retrieval processor from some type of remote input-output device. This is largely a consequence of the explicit research orientation of the Center towards man-machine systems problems in general, and problems of automatic natural language processing in particular.

The basic retrieval processor, which has been described elsewhere [1, 2], is an associative model using index word associations computed automatically by techniques first developed by R. M. Needham [3]. It is being implemented on a small computing system used at The University of Texas for experimental investigation of problems associated with the development and use of time-shared computing. The hardware configuration consists of the elements shown in Figure 1.

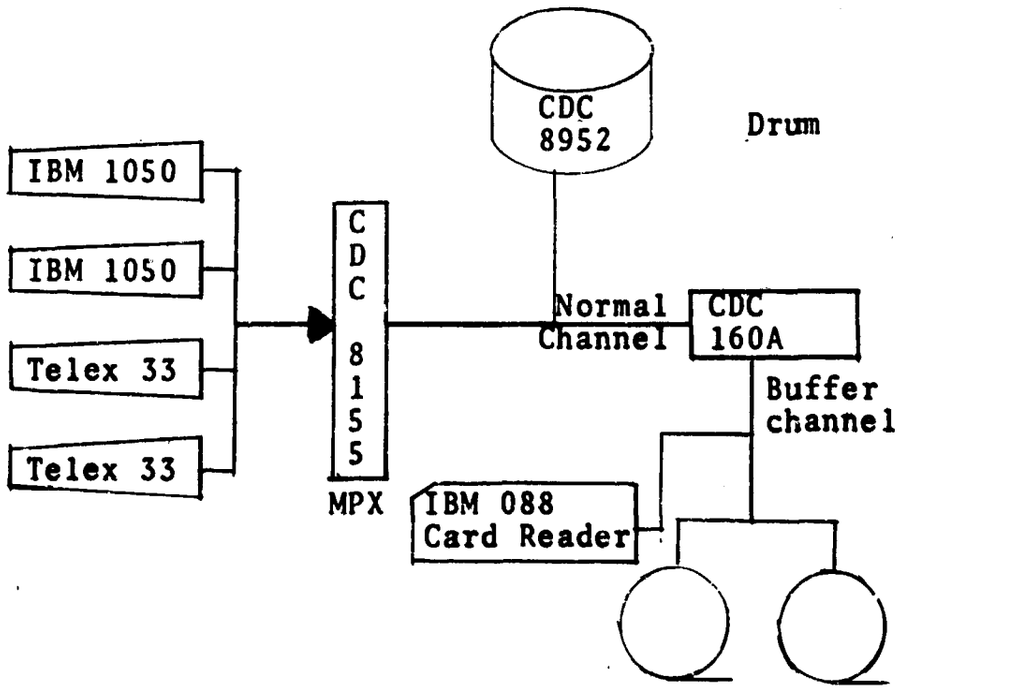
This type of environment permits the development of systems with a number of interesting properties, notably (a) direct file interrogation without interposition of an intermediary, and (b) dialogue between the requester and the retrieval processor, so that retrieval need not be a one-pass operation.

Accordingly, we are planning to extend the present retrieval model to incorporate feedback processes that will permit the requester to refine his search specifications on the basis of inspection of successive outputs from the re-

trieval processor. The initial request to the system is in the form of a list of index words and initial output is a requested (variable) number of documents from the top of a list ordered by a relevancy-scoring algorithm. Subsequent search specifications are in the form of requests to retrieve or avoid documents similar to specified documents previously displayed by the system and examined by the requester. Each iteration produces an ordered list, from which a variable number of items can be selected for display and examination.

Figure 1

Computing System for Retrieval Experiments



Hardware characteristics:

160A central processor:	8192 words (12 bits); 6.4 μs cycle time
8952 magnetic drum:	65,536 words (12 bits); 23 ms average block access time
606 tape units:	30 kc transfer rate
8155 multiplex unit:	16 full-duplex communication channels; generates I/O interrupts every 90 ms.
IBM 1050:	typewriter-like keyboard and carriage. Modified to transmit/receive maximum 7.5 characters per second rather than the usual 15.
Telex 33:	standard Model 33 teletype machine.

2 SYSTEM EVALUATION

In considering the problem of evaluating the performance of this type of system it is felt that criteria other than the Cleverdon-defined relevance and recall ratios should be employed. As Doyle pointed out some time ago [4]:

> In visualizing systems of this kind we can feel the usefulness of the concept of relevance slipping through our fingers. We now become aware that the "most relevant subset" is not only an individual matter for the searcher, dependent on the time and circumstances of his searching foray, but also that the feedback he gets is quite capable of changing his way of expression. An "information need" is thus revealed to be a dynamic entity, whose times of greatest dynamism and change may come in the very process of interacting with a retrieval system.

Additionally, as O'Connor [5], amongst others, has observed, distinctions must be made in considering types of search requests:

> Does the user want any one S-document (to answer a question), a few (to start on a subject), most in the collection (for a good grasp of the subject), or all in the collection (an exhaustiveness needed for scientific, military, safety, or legal purposes)?

These considerations suggest that in examining systems of the type described, we must (a) categorize searches, perhaps along the lines suggested by O'Connor, and (b) use evidence from user behavior as data for evaluation purposes.

At the present time we are particularly concerned with the relationship between indexing procedures and the vocabulary classifications produced by the automatic classification algorithms. Specifically, we wish to investigate the efficiency of retrieval with two basic models:

1. Classification of the entire vocabulary, (i.e. a full associative model).

2. Classification of a subset of the vocabulary, using associations within the subset in conjunction with term coordination for infrequently used index words.

The experimental procedure planned is as follows:

1. Classify each retrieval search according to type.

2. Users will make their own searches in the manner described using the full associative model, the data base for which is a 2000 document collection indexed with a vocabulary of 800 words, with an average of 16 index words per document. Statistics maintained by the system will include:

   (a) Initial search specification

   (b) Documents retrieved

   (c) Documents accepted by the user

   (d) Documents rejected by the user

In general, the full protocol of the user during a given search will be recorded during each search iteration. Users will be required to continue a search until they are satisfied that their information need has been met.

3. The collection will be re-indexed using subsets of the total indexing vocabulary, and reclassification of the modified vocabularies will be made. Existing programming systems permit this entire process to be accomplished automatically [6].

4. Using data on initial search requests, document acceptances and rejections, and number of documents requested at each iteration, previous search patterns will be simulated in the new environment, using previous search protocols on acceptances and rejections.

5. We wish to test the hypothesis that retrieval efficiency has been improved by a given reindexing scheme. The desired performance criterion is rapidity of convergence on a set of accepted documents. One simple proposed scoring polynomial for rating search efficiency is as follows:

A_j = no. of accepted documents on the jth search iteration

N = total number of documents scanned through the last iteration on which a document was accepted (the nth iteration)

S, the measure of search efficiency, may be computed as:

$$S = \frac{\sum_{j=1}^{n} A_j / j}{N}$$

A function such as this has the useful property of giving higher values to searches in which the density of accepted documents is greater in the earlier iterations, in addition to giving weight to a specific acceptance ratio.

Assuming that there are m searches of a given type, there will be m pairs of search scores, S and S' for the two indexing schemes being compared. If $D_i = S_i - S_i'$; $\overline{D} = \sum_{i=1}^{m} D_i$; and $s_{\overline{D}} = s_D / \sqrt{m}$, where s_D is the standard deviation of the m values of D, then $\overline{D}/s_{\overline{D}}$ is distributed as t with m-1 degrees of freedom [7]. We can thus test the hypothesis that the variable D has significantly changed -- that is, that there has been a significant improvement in the set of search scores.

3 SUMMARY

In place of previously used absolute parameters of retrieval performance, we propose a somewhat weaker measure of relative efficiency, appropriate to the particular type of retrieval system under investigation at the Linguistics Research Center, and to the particular problem of investigating the effects of controlled changes in indexing techniques within the system. It appears that desirable features also include:

(a) Classification of types of request and evaluation of system performance separately with reference to each type.

(b) The use of simulation techniques to permit rapid generation of experimental statistics.

REFERENCES

1. A. G. Dale and N. Dale, "Some Clumping Experiments for Associative Document Retrieval," American Documentation, Vol. 16, No. 1, January 1965.

2. A. G. Dale and N. Dale "Clumping Techniques and Associative Retrieval," NBS-ADI Symposium on Statistical Association Methods for Mechanized Documentation, March 1964.

3. R. M. Needham, The Theory of Clumps II, Cambridge, England: Cambridge Language Research Unit, 1961.

4. Lauren B. Doyle, "Is Relevance an Adequate Criterion in Retrieval System Evaluation?" Automation and Scientific Communication, October 1963.

5. John O'Connor, "Mechanized Indexing Methods and Their Testing," Journal of the ACM, Vol. II, No. 4, October 1964.

6. Richard Jernigan, "CENI: A Document Processing Program," LRC 65 WT-1, Austin: Linguistics Research Center, November 1965.

7. Edward C. Bryant, Statistical Analysis, New York, 1960.